Data Cleansing

เปลี่ยน ‘ข้อมูลสกปรก’ ให้เป็น “ขุมทรัพย์เชิงกลยุทธ์”

คู่มือฉบับสมบูรณ์สู่การสร้าง Data Health ที่ยั่งยืน

มายาคติของ ‘Big Data’

ปริมาณ ≠ คุณภาพ
(Volume ≠ Value)

‘ข้อมูลสกปรก’ (Dirty Data)
‘ข้อมูลลกมถูกต้อ Coral Red
ที่ขาดความถูกต้อง ครบถ้วน หรือเป็นปัจจุบัน
นำไปสู่การตัดสินใจเชิงกลยุทธ์ที่ผิดพลาด

หัวใจสำคัญของการใช้ข้อมูลให้เกิดประโยชน์สูงสุด
ไม่ได้อยู่ที่ ‘ปริมาณ’ แต่อยู่ที่ ‘คุณภาพ’

ปรัชญาหลักของ Data Cleansing

Quality Input

(ข้อมูลนำเข้าคุณภาพสูง)

Quality Insight

(ผลการวิเคราะห์คุณภาพสูง)

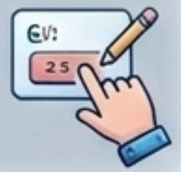
Data Cleansing ไม่ใช่การทิ้งข้อมูลแบบเสียเปล่า
แต่คือการ **‘Re-organize’ จัดระเบียบโครงสร้างใหม่**
เพื่อกำจัดข้อมูลซ้ำซ้อนและไม่เกี่ยวข้อง ลดภาระระบบ และดึง Insight ที่ทรงพลังที่สุด

ทำไมองค์กรต้องให้ความสำคัญกับ Data Cleansing?

กำจัดข้อผิดพลาด:
ลดความคลาดเคลื่อนจากแหล่งข้อมูลที่หลากหลาย (Multiple Sources)

จับคู่เทคนิควิเคราะห์:
จัดรูปแบบให้พร้อมใช้งานและประมวลผลเร็วขึ้น

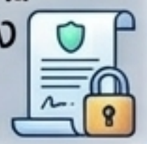
ปรับปรุงกระบวนการ:
มองเห็นรูปแบบ Human Error เพื่อแก้ไขตั้งแต่ต้นทาง



เพิ่มความแม่นยำ:
สร้าง Report และ Insight ที่น่าเชื่อถือ นำไปสู่การตัดสินใจที่ถูกต้อง

เพิ่มความคล่องตัว:
ข้อมูลพร้อมใช้ทันที ลดเวลาการเตรียมข้อมูล

ความเสี่ยงทางกฎหมาย (PDPA):
จัดการข้อมูลที่หมดอายุหรือไม่มีความจำเป็นเพื่อปฏิบัติตามกฎหมายคุ้มครองข้อมูลส่วนบุคคล

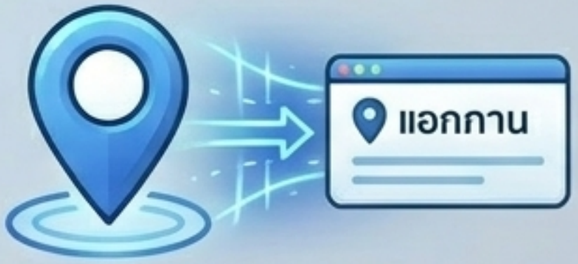


ตรวจสอบอาการ: ข้อมูลแบบไหนที่ต้องนำมาทำ Cleansing?

1. รูปแบบ/ไฟล์ต่างกัน

มาจากหลายแหล่ง
(Multiple Sources)
เช่น PDF, Excel, CSV
ไม่สามารถประมวลผลร่วมกันได้
ต้องทำ Data Transformation

2. โครงสร้างไม่เหมาะสม



ข้อมูลมีอยู่แต่ใช้วิเคราะห์
ไม่ได้ทันที
เช่น พิกัดละติจูด-ลองติจูด
ต้องแปลงเป็นเขต/อำเภอ
(Data Structuring)

3. ไม่ถูกต้อง / ไม่ครบถ้วน

เกิดจาก
Human Input Error
(สะกดผิด) หรือระบบ
(Data Latency, บันทึก
Event Tracking ไม่ครบ)

The Data Refinery Pipeline: กระบวนการทำ Data Cleansing 5 ขั้นตอน

Step 1: กรองข้อมูล (Filtering)

กำจัดข้อมูลที่ซ้ำซ้อน (Removing Duplicates)
และข้อมูลที่ไม่เกี่ยวข้อง (Irrelevant Data)

ลดความซ้ำซ้อน:

ป้องกันผลวิเคราะห์คลาดเคลื่อนจากการบันทึกซ้ำซ้อนจากหลายแหล่ง

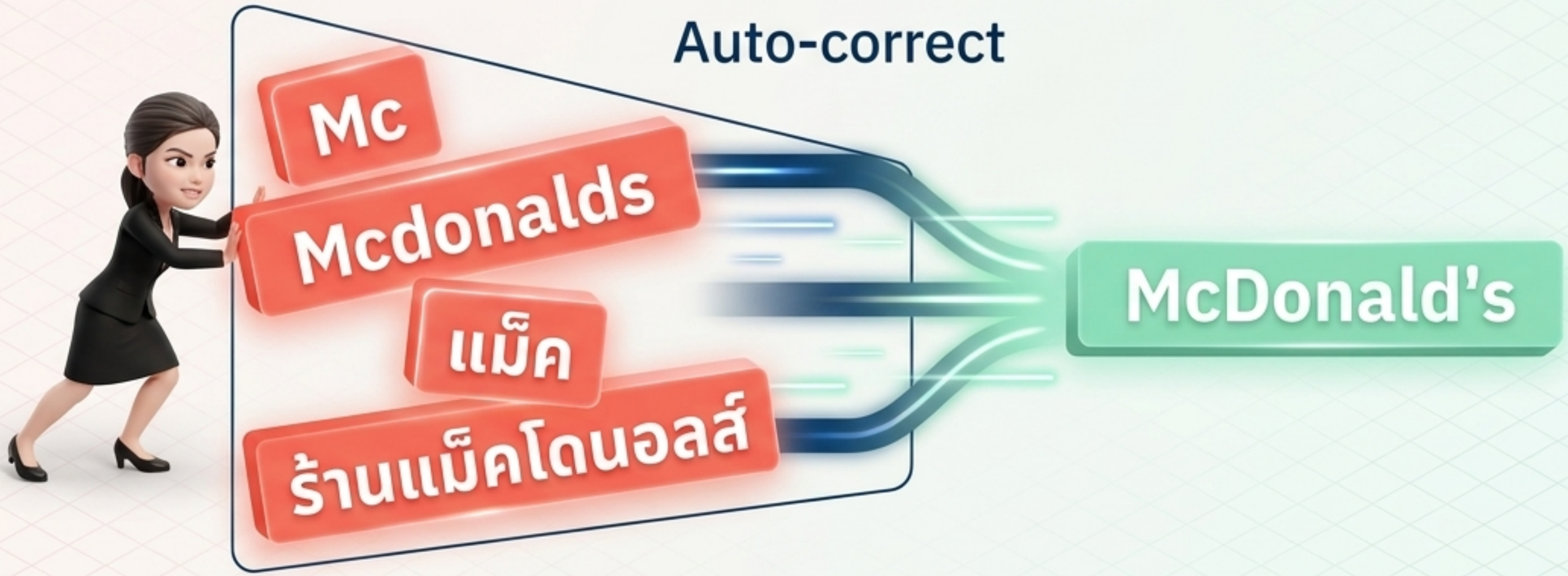
ตัดข้อมูลไม่จำเป็น:

หากต้องการวิเคราะห์ยอดขาย ไม่ต้องใช้ข้อมูลส่วนตัวลูกค้าที่ไม่เกี่ยวข้อง

ลด Processing Load & เพิ่มประสิทธิภาพ

Step 2: ปรับมาตรฐาน (Standardization)

แก้ไขข้อผิดพลาดเชิงโครงสร้าง (Correcting Structural Errors)

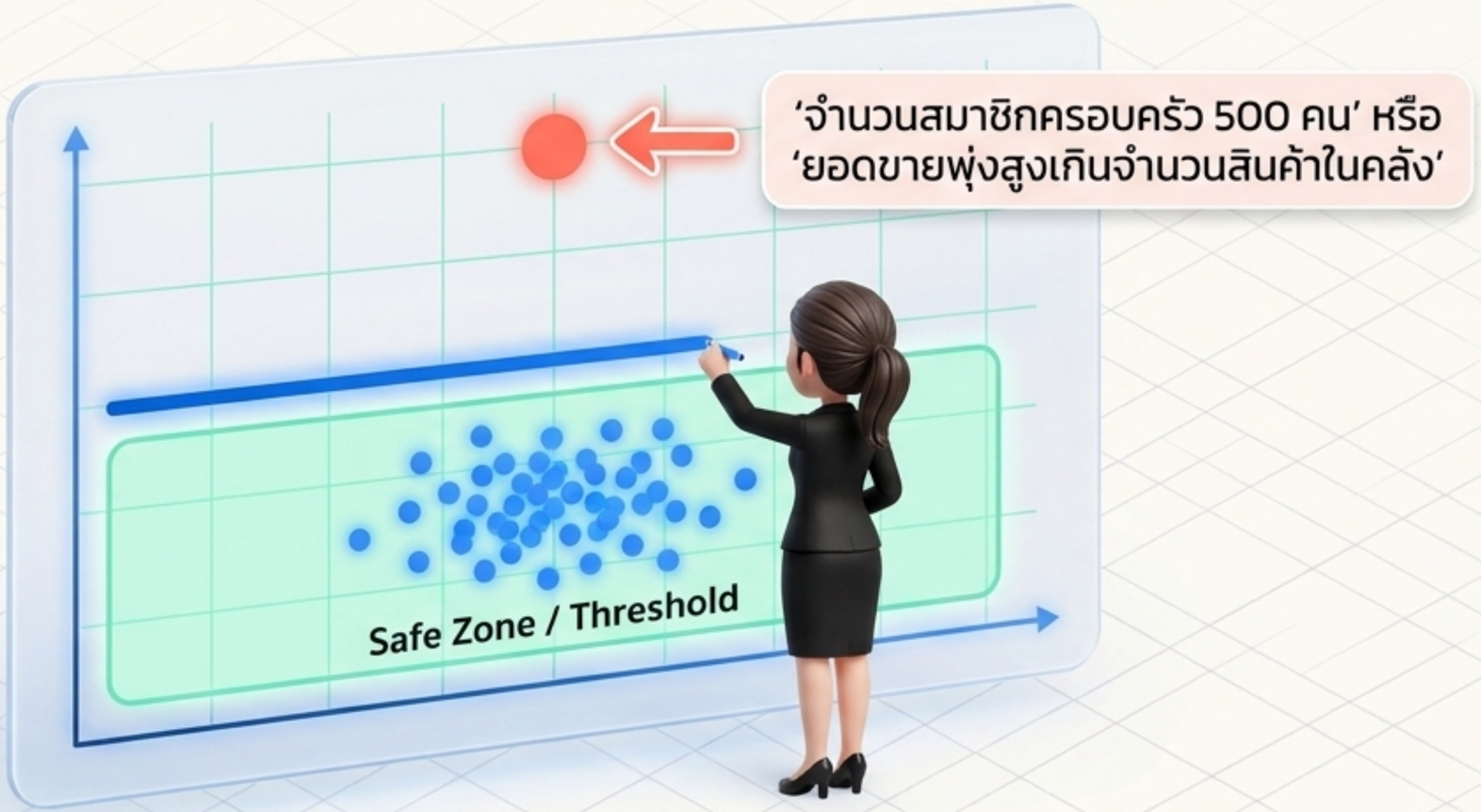


ปัญหาใหญ่ที่ทำให้ Insight ของธุรกิจ ‘เพี้ยน’
คือความไม่สอดคล้อง (Human Error)
หากไม่ทำ Cleansing ระบบจะเข้าใจว่าเป็นคนละร้านกัน

ปรับข้อมูลที่มีความหมายเดียวกันให้อยู่
ในรูปแบบเดียวกัน รวมถึงหน่วยและตัวเลข

Step 3: จัดการค่าผิดปกติ (Handling Outliers)

ตรวจจับข้อมูลที่บิดเบือนความเป็นจริง



ค่าเหล่านี้ (Outliers) หากปล่อยไว้จะบิดเบือนค่าเฉลี่ยและทำให้การพยากรณ์ทางธุรกิจคลาดเคลื่อน ต้องกำหนด **‘เส้นกั้น’** (Threshold) เพื่อกรอง Noise ออกไป

Step 4: กลยุทธ์จัดการข้อมูลสูญหาย (Missing Data)

เลือกวิธีจัดการเมื่อข้อมูลไม่สมบูรณ์

1. ลบทิ้ง (Deletion)

ลบข้อมูลที่ไม่สมบูรณ์ออก
(ต้องระวังผลกระทบต่อภาพรวมการวิเคราะห์)

2. เติมค่า (Imputation)

$\frac{\Sigma}{n}$

แทนค่าด้วยหลักการ เช่น ใช้ค่าเฉลี่ย (Mean)
ค่ากลาง (Median) หรือการประมาณค่า

3. เปลี่ยนวิธี (Change Method)

ปรับเปลี่ยนวิธีการวิเคราะห์
หรือเลือกใช้ชุดข้อมูลอื่นที่เหมาะสม
กับวัตถุประสงค์มากกว่า

Step 5: ตรวจสอบความถูกต้อง (Data Validation / QA)

ด่านสุดท้ายก่อนนำไปสร้าง Insight

- ข้อมูลมีความสมเหตุสมผลหรือไม่? (Reasonableness)
- รูปแบบข้อมูลสอดคล้องกับวัตถุประสงค์หรือไม่? (Alignment)
- ข้อมูลสามารถนำไปใช้สร้าง Insight ได้จริงหรือไม่? (Actionability)

หากพบข้อบกพร่อง ต้องย้อนกลับไปปรับปรุงข้อมูลในขั้นตอนก่อนหน้าอีกครั้ง

อนาคตขององค์กรอยู่ที่ ‘สุขภาพของข้อมูล’ (Data Health)

องค์กรที่ประสบความสำเร็จ ไม่ใช่องค์กรที่มี
‘ข้อมูลจำนวนมาก’ เพียงอย่างเดียว
แต่คือองค์กรที่บริหารจัดการ ‘คุณภาพของข้อมูล’
ได้อย่างมีประสิทธิภาพ

ความแตกต่างระหว่างผู้นำและผู้แพ้ในยุคดิจิทัล
คือความสามารถในการทำ **Data Cleansing**
เพื่อสร้างรากฐานข้อมูลที่ถูกต้อง ครบถ้วน และเชื่อถือได้
นำไปสู่ความได้เปรียบในการแข่งขันอย่างยั่งยืน